DOCUMENT RESUME

ED 219 414 TM 820 425

AUTHOR Huberty, Carl J.
TITLE Some Univariate-Multivariate Generalizations.
PUB DATE Mar 82
NOTE 28p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Analysis of Variance; *Data Analysis; Hypothesis Testing; Matrices; *Multivariate Analysis; *Statistical Analysis; Statistics

ABSTRACT

The mathematical extension from scalars to matrices as a prerequisite for univariate to multivariate generalizations of statistical concepts and methods is discussed. One reason for several alternative techniques is the multitude of ways to generalize from univariate concepts to multivariate concepts. The additional complexity, and lack of theoretical and empirical evidence to support a best method, are two other reasons. The basic notion of variability is generalized from a univariate context to a multivariate context using two matrix functions, a determinant and a trace. These two methods of generalizing variance are shown to yield several alternative multivariate indices of shared variation. Multivariate partial indices are proposed. The complexity of multivariate test statistics and some problems in the interpretation of tests of multivariate hypotheses are reviewed. (Author/DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED219414

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Some Univariate-Multivariate Generalizations

Carl J Huberty
University of Georgia

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

C. J. Huberty

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the Annual Meeting of the American Educational Research Association, New York, March 1982.

TM 820 425

Abstract

The mathematical extension from scalars to matrices as a prerequisite for univariate to multivariate generalizations of statistical concepts and methods is discussed. The basic notion of variability is generalized from a univariate context to a multivariate context using two matrix functions, a determinant and a trace. These two means of generalizing variance are shown to yield a number of alternative multivariate indices of shared variation. Multivariate partial indices are proposed. The complexity of multivariate test statistics and some problems in the interpretation of tests of multivariate hypotheses are reviewed.

Some Univariate-Multivariate

Generalizations

Scalars and Matrices	1	
Multivariate Variability		1
Multivariate Sums of Products		3
Shared Variation	4	
Multivariate Indices		5
Choosing an Index		7
Partial Indices		11
Significance Testing	12	
Canonical Correlation		12
MANOVA and Canonical Correlation		14
Shared Variation		16
Discussion	17	

Some Univariate-Multivariate

Generalizations

Scalars and Matrices

Statistical theory and methods pertaining to one variable, and to some extent two variables, typically involve simple numbers, or scalars. Some common scalar symbols are μ, σ^2, ρ_{12}, $\bar{Y}$, s^2, r_{12}, and SSE. When dealing with the theory and calculations related to multivariate statistics and analysis, however, collections of numbers, vectors or matrices, play a vital role. The arithmetic/algebra of vectors and matrices provides a conciseness of formulation in transmitting the complexities of multivariate statistical theory, and in presenting convenient calculation algorithms in multivariate analysis. There are mean vectors (e.g., $\underline{\bar{Y}}$, $\underline{\mu}$), vectors of weights (e.g., $\underline{b}$, $\underline{\beta}$), correlation matrices (e.g., R), covariance matrices (e.g., S, Σ), and sums-of-squares and cross-products (SSCP) matrices (e.g., E, H, T), to name a few.

Multivariate Variability

The generalization of some scalars to matrices is fairly straightforward in a notational sense: e.g., $\bar{Y} \rightarrow \underline{\bar{Y}}$, $\mu \rightarrow \underline{\mu}$, $s^2 \rightarrow S$, $\sigma^2 \rightarrow \Sigma$, and SSE $\rightarrow$ E. Certain problems arise, however, when it comes to assessing and describing multivariate distributions using convenient single numbers rather than collections of numbers. For example, even though an obvious multivariate generalization of a sample variance, s^2, is the sample covariance matrix, S, it is more convenient to have a single number as an index of multivariate scatter. An index that is often proposed is the sample generalized variance, the determinant of S, $|S|$. [The geometric rationale

for this index of variability of a set of variables is given by Tatsuoka (1971, pp. 84-85) and Green (1976, pp. 122-124).] Large values of this index reflect a high degree of scatter about the mean vector, $\underline{\bar{Y}}$, low values reflect concentration about $\underline{\bar{Y}}$. This index plays an important role in maximum likelihood estimation, and "shows its face" often in multivariate-normal theory and methods.

It should be pointed out that the generalized variance index reflects but one aspect of variability in multivariate data. That there is a second aspect may be seen by examining a statistic used in testing H_o: $\Sigma = \Sigma_o$; the statistic involves not only $|S|$ but the trace of S, tr(S), as well (see Morrison, 1976, p. 248):

$$\nu \{\ln|\Sigma_o| - \ln|S| + \mathrm{tr}(S \cdot \Sigma_o^{-1}) - p\},$$

where ν = df for S, and p = number of variables. [The trace of S is the sum of the main diagonal elements of S.] The trace of S reflects the second aspect, that of total variation, i.e., an aggregate of variances. Total variation is a useful concept in principal components analysis, where a goal is to summarize most of the variability in the data using only the principal components with the highest variances. [It may be noted that tr(S) is, in fact, a generalization of variance in the univariate case in which tr(S) = s^2.]

As an example of how using only $|S|$ as an index of scatter may not tell the whole story, consider the two bivariate covariance matrices:

$$S_1 = \begin{bmatrix} 8 & 8 \\ 8 & 9 \end{bmatrix} \quad \text{and} \quad S_2 = \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix} .$$

Now $|S_1| = |S_2| = 8$, whereas $\mathrm{tr}(S_1) = 17$ and $\mathrm{tr}(S_2) = 6$. It might be

concluded that there is identical scatter of the data vectors about the two-element mean vectors. There is, however, greater total variation in the first data set. The latter relative assessment may only be made, of course, when the variable measures in the two sets are comparably scaled.

The fact that there are two different aspects of multivariate variability has some implications for data simulation in multivariate methodological studies. In specifying covariance structure for a "population" from which sampling is to be done, the trace as well as the determinant of a population covariance matrix might well be considered. By so doing, one can more completely describe and specify the populations being studied.

Multivariate Sums of Products

In generalizing from univariate analysis of variance (ANOVA) to multivariate analysis of variance (MANOVA), there are some "natural" scalar-matrix counterparts. An ANOVA hypothesis sum of squares, SSH, generalizes to H, an hypothesis SSCP matrix. Similarly, an "error" sum of squares generalizes to E, and a total sum of squares to T. Just as in the case of arriving at a multivariate index of variation, it is desirable to consider single numbers as multivariate counterparts to univariate sums of squares. As will be seen in the next section, and following the generalized variance idea, it is useful to consider determinants of matrices as the single numbers. We thus have the following possible generalizations:

SSH ——————> $|H|$,

SSE ——————> $|E|$, and

SST ——————> $|T|$.

There is a problem with the first generalization, however, since $|H|$ is usually zero. That this is the case may be seen by considering the rank H, or equivalently, its associated df value. For a one-way layout, the rank

of the p x p matrix H is the minimum of p, the number of variables, and k - 1, where k is the number of "groups." Since p is usually greater than k, $|H|$ is usually zero.

Shared Variation

The square of the Pearson correlation coefficient, or the coefficient of determination, r^2, expresses the proportion of the variance "shared" by the two variables (when the two variables are linearly related). In a multiple linear correlation context, R^2 expresses the proportion of variance of one variable, given by an "optimum" linear composite of the other variables. As intended here, these two contexts of shared variation involve a single group of experimental units on each of which are available the measures of interest. The notion of shared variation has also been applied in the multiple-group context. In this context are found the popular indices η^2 and ω^2. Strictly in terms of estimation, the statistics η^2 and ω^2 have different properties. For descriptive purposes, however, both may be roughly interpreted as the proportion of variance in the response variable measures that may be attributed to the effect associated with the hypothesis being tested. In a one-way layout, an η^2 or ω^2 value would reflect in a descriptive sense the proportion of criterion variance attributable to the single grouping variable. In a two-way layout, an η^2 or ω^2 value associated with the interaction effect, for example, would reflect the proportion of criterion variance attributable to the joint effect of the two grouping variables.

Just as when interpreting an r^2 value or an R^2 value, care must be taken in interpreting an η^2 or ω^2 value. One "care" to which reference is made deals with how the variables are defined/measured. For example, when discussing variation shared by "method of instruction" and "intelligence,"

interpretation is highly dependent on how the methods are defined and how intelligence is measured.

For an excellent discussion of univariate indices of shared variation see Maxwell, Camp, and Arvey (1981).

Multivariate Indices

Various proposals have been made for generalizing the notion of shared variation, or strength of association, to multivariate situations beyond those involving R^2 values. Cramer and Nicewander (1979) provide an excellent review and discussion of shared variation indices in a single-group, canonical correlation context. They present seven indices, three of which are original, and all of which are generalizations of R^2 in that all reduce to R^2 in the case where one set of variables is composed of a single variable. One of their original indices, of interest here, is the average squared canonical correlation. [See, also, Srikantan, 1970.]

In the multiple-group context, too, several generalized indices of shared variation have been proposed. Since the basic concept in shared variation is that of variance, it is not too surprising that the two transformations of matrices discussed in the last section appear in the generalizations. Focusing on generalized variance, determinants of SSCP matrices were used by Wilks (1932), Tatsuoka (1970), Smith (1972), and Sachdeva (1973). Focusing on total variance, traces of products of SSCP matrices were used by Hsu (1940), Stevens (1972), Coxhead (1974), Shaffer and Gillo (1974), and Serlin (1981).

In expressions for the indices, the following definitions are given: p is the number of response variables; E, H, and T are the p x p error, hypothesis, and total SSCP matrices, respectively; df_e is the error df value; df_h is the hypothesis df value; the Hotelling-Lawley trace statistic

is $V = tr(E^{-1}H)$; and $r = \min(p, df_h)$. Although nine papers were cited above, only six indices are taken from the papers since (i) Tatsuoka and Sachdeva independently arrived at the same index, (ii) the Smith formulas, based on stepdown procedures, "do not yield values that are invariant under alternate orderings" (Smith, 1972, p. 371) of the response variables, a property desirable in situations other than those of interest here, and (iii) Coxhead and Shaffer and Gillo independently arrived at the same index.

The Wilks index is simply

$$I_1 = 1 - \Lambda, \tag{1}$$

where $\Lambda = |E| \; / \; |E + H|$. The Tatsuoka-Sachdeva index, for the one-way layout situation with k groups and a total of N units, is given by

$$I_2 = \frac{|T| - |E| - \frac{df_h}{df_e}|E|}{|T| + \frac{1}{df_e}|E|} \tag{2}$$

or

$$I_2 = 1 - \frac{N \Lambda}{df_e + \Lambda} \; . \tag{2'}$$

The Hsu index is

$$I_3 = \frac{V}{1 + V} \; . \tag{3}$$

The Stevens index is a modifcation of (3):

$$I_4 = \frac{V'}{1 + V'}, \tag{4}$$

where $V' = (df_e - p - 1) \; V \; / \; df_e$. The Coxhead-Shaffer-Gillo index is

$$I_5 = \frac{V}{p + V} \; . \tag{5}$$

The Serlin index, based on the Cramer-Nicewander notion of an average squared

canonical correlation, is

$$I_6 = U/r \ , \tag{6}$$

where $U = tr\ [(H + E)^{-1} H]$ is the Bartlett-Pillai trace statistic. The last index to be considered is one also based on traces of SSCP matrices:

$$I_7 = \frac{tr(H)}{tr(H+E)} \ . \tag{7}$$

Choosing an Index

With the seven currently available descriptive indices, and perhaps more to come in the future, how does a researcher choose one to use? The choice is not simple. For a one-way layout, all of the indices except I_2 are equivalent when $p = 1$ to the univariate index,

$$\eta^2 = \frac{SSH}{SSH + SSE} = 1 - \frac{SSE}{SSH + SSE} \ . \tag{8}$$

With $p = 1$, this is the case for I_3, I_4, and I_5 since $V = SSH/SSE$, for I_6 since $U = SSH/(SSH + SSE)$, and for I_7 since $tr(H) = SSH$ and $tr(H + E) = SSH + SSE$. [Hsu's index, I_3, may also be viewed as a direct generalization of Cohen's (1977, p. 281) expression for η^2.] The Tatsuoka-Sachdeva index, I_2, simplifies to the univariate ω^2 index,

$$\omega^2 = \frac{SSH - df_h.MSE}{SST + MSE} = \frac{SST - SSE - \frac{df_h}{df_e} SSE}{SST + \frac{1}{df_e} SSE} \ . \tag{9}$$

So, one criterion may be based on a researcher's preference in the univariate case. If preference is for an ω^2-type index, then I_2 would obviously be the choice. If, however, an η^2-type index is preferred, then the researcher must choose one of the other six indices. Choice among these six may be based on a preference for multivariate scatter reflected by a

matrix determinant or by a matrix trace. If a generalized variance notion is preferred (which involves a determinant), then I_1 would be the choice, otherwise one must choose from among I_3, I_4, I_5, I_6, and I_7 that are based on a total variance notion (which involves a trace). In situations where df_e/p or N/p is "large," there is not much to choose between I_3 and I_4. [In fact, if numerical value is a basis of choice, choosing between I_1, I_2, and I_3 could be accomplished by drawing straws (see Huberty, 1972).]

Choosing from among I_3, I_5, I_6, and I_7 presents, in general, some conceptual problems. [For the particular multivariate situation of two groups (i.e., $k = 2$ and $r = 1$), $I_1 = I_6$.] The rationale behind the Coxhead-Shaffer-Gillo index, I_5, is appealing and "has a clear interpretation in terms of discriminant functions," in that "it can be regarded as a weighted average of the estimated correlation ratios for each of the discriminant functions" (Shaffer & Gillo, 1974, p. 523). The rationale behind the Serlin index, I_6, is also appealing since it is a mean of the squared canonical correlations between the set of response variables on one hand, and a set of dummy variables on the other.

Ease of calculation is certainly no basis for choice. With standard computer program output, values of all six indices are relatively easy to calculate. The value of Λ is typically given; thus values of I_1 and I_2 are easily determined. To find values of I_3, I_4, and I_5 it may be noted that $V = tr(E^{-1}H) = \sum_{i=1}^{r} \lambda_i$, where λ_i is the ith eigenvalue of $E^{-1}H$; values of λ_i are also typically given. Since $U = \sum_{i=1}^{r} \frac{\lambda_i}{1 + \lambda_i}$, a value of I_6 can be determined. [An I_6 value is yielded by the STEPDISC Procedure of SAS's (1981) 79.5 Release; SAS's GLM Procedure explicitly outputs values of V and U as well as that of Λ.] A value of I_7 is obtainable by using SAS's GLM program to print the E and H matrices.

A final criterion of choice which may satisfy some researchers is based on the idea of parameter estimation. Tatsuoka (1973) has pointed out that, with the current notation, I_1, I_2, and I_3 "are all equally poor instead of equally good" (p. 31). This statement was made in reference to using these indices as estimators of true shared variation. Tatsuoka arrived at an adjusted or "shrunken" multivariate ω^2:

$$\tilde{I}_2 = I_2 - \frac{p^2 + (k - 1)^2}{3N} (1 - I_2) .$$

About 60 years ago K. Pearson suggested a less biased estimator of the population correlation ratio in a univariate sense (see Carroll & Nordholm, 1975). Generalizing Pearson's adjustment to the multivariate context, the validity of which remains to be shown, yields, for $k > 3$,

$$\frac{I_1 - (k - 3)/N}{1 - (k - 3)/N} .$$

Tatsuoka (1973, p. 31) conjectured that the adjustment used with I_2 will "suffice for all practical situations" when used with I_1 and I_3. If so, we may use

$$\tilde{I}_1 = I_1 - \frac{p^2 + (k - 1)^2}{3N} (1 - I_1) ,$$

and

$$\tilde{I}_3 = I_3 - \frac{p^2 + (k - 1)^2}{3N} (1 - I_3) .$$

Serlin (1981) has also proposed an adjusted I_6 index, employing the idea of shrinkage used by M. Ezekiel in a multiple correlation context (see Huberty & Mourad, 1980, p. 103):

$$\tilde{I}_6 = I_6 - \frac{b}{N - b - 1} (1 - I_6) ,$$

where $b = \max(p, k - 1)$. Adjustments for I_4, I_5 and I_7 have not been proffered.

As reflective as these comments may be, they do not provide much guidance for choosing a single preferred index. Even though Cramer and Nicewander (1979) would prefer I_6 to I_5, they imply that a single index will not provide satisfactory information regarding shared variation. Comparisons of the generalized descriptive indices are conceptually difficult and, perhaps, inappropriate. Different interpretations of shared variation are reflected across the indices.

As an illustration of the numerical comparability of the seven index values, real data in a one-way layout were analyzed with $p = 15$ response variables, $k = 3$ groups, and $N = 200$. The seven values are:

$$I_1 = 1 - .591 = .409,$$

$$I_2 = 1 - \frac{200(.591)}{197 + .591} = .402,$$

$$I_3 = \frac{.634}{1 + .634} = .388,$$

$$I_4 = \frac{.582}{1 + .582} = .368,$$

$$I_5 = \frac{.634}{15 + .634} = .041,$$

$$I_6 = \frac{.444}{2} = .222, \text{ and}$$

$$I_7 = \frac{12416.216}{417092.540} = .030.$$

Proposed adjustments yield $\tilde{I}_1 = .184$, $\tilde{I}_2 = .174$, $\tilde{I}_3 = .154$, and $\tilde{I}_6 = .159$.

The numerical ordering of the six indices obtained in the above example is not that to be expected generally. An ordering of one subset of the set of six indices is, however, obvious: $I_5 \leq I_4 \leq I_3$. Another index based

on a trace, I_6, may be within or outside of this inequality. The two indices based on a determinant, I_1 and I_2, generally have values comparable to those of I_3 and I_4.

Partial Indices

In a univariate two-way layout context, "partial" (actually, semi-partial) indices have been proposed by Cohen (1973) for η^2 and by Keren and Lewis (1979) for ω^2. For a given hypothesis, a partial η^2 may be defined by

$$\eta_h^2 = \frac{SSH}{SSH + SSE} = 1 - \frac{SSE}{SSH + SSE} . \tag{10}$$

A partial ω^2 may be defined by

$$\omega_h^2 = \frac{(SSE + SSH) - SSE - \frac{df_h}{df_e} SSE}{(SSE + SSH) + \frac{N - df_e - df_h}{df_e} SSE} . \tag{11}$$

[Note the resemblances between (8) and (10) and between (9) and (11).]

There is a "natural" generalization from the univariate partial η^2 to a multivariate partial η^2:

$$I_1' = 1 - \Lambda_h, \tag{12}$$

where $\Lambda_h = |E| / |E + H|$ is the Wilks lambda statistic for a given hypothesis. One "slicing of the matrix pie," to borrow a phrase from M. M. Tatsuoka, yields a multivariate partial ω^2

$$I_2' = \frac{|E + H| - |E| - \frac{df_h}{df_e} |E|}{|E + H| + \frac{N - df_e - df_h}{df_e} |E|} . \tag{13}$$

Tatsuoka (1970, 1973) has poined out that ω^2 may be expressed in terms of the Wilks lambda statistic [see (2')]. Similarly, (13) may be expressed as

$$I_2' = 1 - \frac{N \Lambda_h}{df_e + \Lambda_h (N - df_e - df_h)} . \tag{13'}$$

Partial indices corresponding to I_3, I_4, I_5, and I_6 are obviously obtainable considering the definitions of V and U. A value of I_7' is obtained from the traces of E and H. The needed numerical values may be found in the output of SAS's GLM Procedure.

As an illustration of the calculation of partial indices, real data in a two-way (2 x 4) layout with p = 6 and N = 148 were analyzed. Values of the partial indices are:

	I_1'	I_2'	I_3'	I_4'	I_5'	I_6'	I_7'
A	.126	.114	.126	.120	.023	.126	.014
B	.205	.183	.198	.190	.039	.071	.068
AB	.113	.091	.110	.105	.020	.039	.020

[Based on significance testing, the AB effects were judged to be nonexistent.]

In a multivariate situation, just as in univariate data analysis, there is the problem of assessing the magnitude of a shared variation value. That is, is an I_1' value of 0.126 "small" or "large"? How about an I_5' value of 0.039 or an I_6' value of 0.071? These are open questions, answers to which are situation- and researcher-dependent.

Significance Testing

Canonical Correlation

It is well known that the concept and formulation associated with simple linear (i.e., Pearson) correlation generalize to those associated with multiple rectilinear correlation which, in turn, generalize to those with canonical correlation. Formulation-wise, this may be seen by examining the determinantal equation typically given in the mathematics of canonical correlation:

$$|R_{yy}^{-1} R_{yx} R_{xx}^{-1} R_{xy} - \theta I| = 0, \qquad (14)$$

where R_{yy}, R_{xx}, and $R_{yx} = R'_{xy}$ are correlation matrices of dimensions p x p, q x q, and p x q, respectively, and θ represents the set of eigenvalues obtained in the solution of the equation. In the multiple correlation context with p = 1, (14) simplifies to

$$|R_{yx} R_{xx}^{-1} R_{xy} - \theta I| = 0 . \qquad (15)$$

But, $R_{yx} R_{xx}^{-1} R_{xy}$ is a scalar (equal to R^2), and the solution of (15) is this scalar; thus, $\theta = R_{yx} R_{xx}^{-1} R_{xy} = R^2$. With p = q = 1, (15) simplifies to $|r_{yx}^2 - \theta| = 0$, or $\theta = r_{yx}^2$.

Another view of this generalization may be made by noting that an R^2 value is simply an r^2 value for a single Y variable versus an optimal linear composite of the q X variables--optimal in the sense of maximum correlation of the Y variable versus the composite. Further, a θ value in (14) is simply an r^2 value for an optimal linear composite of the p Y variables versus an optimal linear composite of the q X variables.

It must be noted that with $p \leq q$, there are p values of θ in the solution set of (14). That is, for $p \geq 2$, there is more than one canonical correlation coefficient (each of which is the positive square root of a θ value). This note is important when significance testing in a canonical correlation or in a MANOVA context is discussed.

The exact statistic used in testing the significance of a Pearson r is a special case of the exact statistic used in testing the significance of a multiple correlation coefficient [see Kerlinger and Pedhazur (1973, pp. 23, 63) for the specific formulas]. It turns out that the first statistic is equivalent to the statistic used in testing the difference between two group means, and the second is equivalent to the statistic used in testing differences among several group means. That is, univariate two-group-mean-

difference hypotheses may be tested in a simple regression/correlation context, and univariate analysis of variance (ANOVA) hypotheses may be tested in a context of multiple regression/correlation. There exists a similar generalization, with an added complexity, of the above relationship to one between multivariate analysis of variance (MANOVA) and canonical regression/correlation.

MANOVA and Canonical Correlation

There is a neat mathematical association between a one-way k-group MANOVA and a canonical correlation analysis (for details, see Tatsuoka, 1971, pp. 183-190). For $r = \min(p, k - 1) = k - 1$, let $\delta_1, \delta_2, \ldots, \delta_r$ denote the set of eigenvalues of the matrix product $(H + E)^{-1} H$. The δ_i values are the squares of the canonical correlation coefficients involving the p variables on the one hand, and $r = k - 1$ dummy variables representing levels of the grouping variable on the other hand. [In the notation of expression (14), $q = k - 1$ and $\delta = \theta$.] It turns out that four test criteria used in testing the null hypothesis for a simple MANOVA,

$$H_o: \underline{\mu}_1 = \underline{\mu}_2 = \ldots = \underline{\mu}_k ,$$

are all functions of $\{\delta_i\}$. Three criteria were mentioned in the immediately preceding section of this paper. The Wilks criterion, Λ, may be expressed as $\Lambda = \prod_{i=1}^{r} (1 - \delta_i)$. The Hotelling-Lawley trace, V, may be expressed as $V = \sum_{i=1}^{r} \lambda_i$, where $\lambda_i = \delta_i/(1 - \delta_i)$ is the ith eigenvalue of $E^{-1} H$. The Bartlett-Pillai trace, U, may be expressed as $U = \sum_{i=1}^{r} \delta_i$. A fourth criterion, due to S. N. Roy, is simply $\delta_1 = \max\{\delta_i\}$, or, equivalently, $\lambda_1 = \max\{\lambda_i\}$.

An equivalent way of stating the null hypothesis is

H_o: The r population canonical coefficients are simultaneously zero. (16)

Two statistics often used in testing H_o are Rao's F statistic (which is approximate unless p = 1 or 2 or k = 2 or 3), and Bartlett's approximate chi-squared statistic, both transformations of Λ (see Tatsuoka, 1971, pp. 86, 88). The special cases of p = 1 or 2 or k = 2 or 3 do not present any difficulties (see Knapp, 1978, p. 412) since for these cases there is an exact F transformation of Λ (Tatsuoka, 1971, p. 89). Approximate transformations of V and of U are also available (see SAS's GLM Procedure). Special tables are needed when δ_1 (or, λ_1) is used as the test statistic, and special tables may be used for Λ, V, and U as well (see, e.g., Timm, 1975, pp. 147-149). A comparative study of the four criteria was reported by Olson (1976).

A comment needs to be made regarding an added complexity in MANOVA and canonical correlation hypothesis testing. Statement (16) suggests a "test of dimensionality" (Mardia, Kent, & Bibby, 1979, p. 338):

H_o: the $\underline{\mu}_j$ lie in an s-dimensional hyperplane,

where $s \leq r = \min(p, k - 1)$. In testing this null hypothesis one is essentially asking the question: What is the minimum number of dimensions (canonical variates or linear discriminant "functions," LDFs) needed to "represent" or account for differences among the $\underline{\mu}_j$? The procedure typically used in answering this question is sequential testing involving a repeated partitioning of Bartlett's approximate chi-squared statistic (see, e.g., Tatsuoka, 1971, p. 165). [Perhaps the reason for typical use is that partitioned chi-squared values are yielded by computer program packages.] This procedure has been severely attacked by Harris (1976) who argued that

statistics claimed to have (approximate) chi-squared distributions do not, and, therefore, are inappropriate for testing hypotheses not only about dimensionality but also about significance of individual canonical variates (or, individual LDFs). That it is inappropriate to use the partitioning for testing significance of individual variates is, at this time, pretty much agreed upon by statisticians. [Harris, (1975, p. 108) suggests a crude conservative test for individual variates.] Mendoza, Markos, and Gonter (1978) have rebutted Harris' argument of the inappropriateness of using the sequential testing procedure for dimensionality, however. It is important for the researcher to appreciate the distinction between two conclusions: (1) intergroup differences may be adequately represented in s = 3 dimensions, and (2) the first, second, and third canonical variates are each significant. As always, it is important to know what hypothesis is being tested by what statistic.[1] See Mendoza et al. (1978) for explicit statements of null hypotheses testable using the sequential testing procedure; rejection of one such hypothesis would lead to the first conclusion, not the second.

Shared Variation

In the univariate two-group problem, testing the significance of the mean difference is equivalent to testing the significance of the point-biserial correlation coefficient. [In fact, it may be shown that the test statistics are algebraically equivalent.] The same may be said for testing mean equality and significance of η^2 in the univariate k-group problem. In this case the test statistics are equivalent in the sense of equivalent referent distributions, F for equality of means and beta for η^2.

In a somewhat similar manner, significance of some of the indices presented in the immediately previous section of this paper may be tested. The significance of $I_1 = 1 - \Lambda$ is determined using one of the approximate

tests for Λ itself. Explicit test statistics for I_2 and I_7 have not been proposed; Tatsuoka (1973), however, empirically arrived at some centile points of the distribution of I_2 for the purpose of constructing interval estimates. Test statistics for I_3 and I_4 may be developed in the manner of Coxhead's (1974, p. 211) development of an approximate F statistic for I_5. The statistics for I_3, I_4, and I_5 will, however, yield identical F values, and will be identical to the approximate F statistic (due to K. C. S. Pillai) yielded by SAS's GLM program for V itself. Serlin (1981) suggested a test statistic for $I_6 = U/r$, but it too is equivalent to that used for U itself (also due to Pillai and given by SAS's GLM program).

Discussion

It has been my experience that some novices in the field of data analysis and statistical methods are disturbed and perplexed when they learn that alternative methods and techniques may be used for analyzing a given set of data. When such a novice expresses his/her disturbance, I

have often been tempted to ask: What is the way to teach the concept of fraction to young children? Or, what is the school administration model to implement? Or, what is the therapeutic method to use with schizophrenics? Expectedly, there are alternatives in data analysis, too. The choice of a method to use will be based on the user's knowledge of, and experience with, various techniques, and his/her general philosophy of data analysis.

The multiplicity of alternative techniques/methods in the realm of univariate data analysis is generally pretty well known. It should come as no surprise, then, that greater freedom of choice (perplexity?) exists in the realm of multivariate data analysis. One reason for the greater freedom of choice is the various ways that one can generalize from univariate concepts and methods to multivariate concepts and methods. [Another reason is, of course, related to the additional complexity; and a third reason is the lack of theoretical and empirical evidence in support of a "best" multivariate method or technique for many situations.] For example, as pointed out in the first section of this paper, the notion of variability, a fairly straightforward univariate notion, can be multivariately interpreted in (at least) two somewhat different ways. These two interpretations, in turn, lead to a number of alternative indices of shared variation. To conclude that one index is "good" and another is "bad" would be a judgment which would require a very specific conceptualization of multivariate variation.

That a multivariate technique/index reduces to a technique/index that is acceptable in a lower-level (e.g., bivariate) data analysis is not a sufficient criterion for its adoption. Probably the overwhelming criterion used in selecting multivariate techniques and indices is the availability

of numerical values on computer printout. More specifically, only those indices are utilized whose explicit values appear on printout. For example, in the near future it can be expected that values of the shared variation index I_6 will be reported in studies because it is explicitly reported in a printout of SAS's recent release.

It is the contention of the current writer that indices of shared variation can and should be utilized in multivariate analyses. They can provide an indication, in addition to a "P-value," of the existence of real effects. And their use need not be restricted to one-way layouts. Researchers have available to them partial indices of shared variation for use in higher-order layouts, some of which are proposed in this paper. In a confirmatory-type study where estimation is of importance, adjusted or "shrunken" index values should be considered; such indices are also proposed in this paper.

Care must be taken when conducting tests of hypotheses in a multivariate context. In particular, what hypotheses are tested with what statistics should be clearly recognized. For example, that a chi-squared statistic often used in multivariate analysis of variance is inappropriate for testing the significance of individual eigenvalues was pointed out in this paper.

Data analysis generalizations from one response variable to two variables to p variables cannot be successfully accomplished in any naive manner. New considerations need to be given with regard to conceptualization, interpretation, and probability distributions. Care must also be given to new formulations. An example of an incorrect generalization involves the calculation of the degrees of freedom value for a three-way contingency table. The df value in a two-way table may be expressed as

$\prod_{i=1}^{2} (c_i - 1)$, where c_i is the number of categories in the ith dimension. Simply replacing the "2" with a "3" is incorrect (see Conover, 1971, p. 165). [The correct df value is $(c_1c_2c_3 - 1) - (c_1 - 1) - (c_2 - 1) - (c_3 - 1) = c_1c_2c_3 - c_1 - c_2 - c_3 + 2$.]

References

Carroll, R.M., & Nordholm, L.A. Sampling characteristics of Kelly's ε^2 and Hays' ω^2. Educational and Psychological Measurement, 1975, 35, 541-554.

Cohen, J. Statistical power analysis for the behavioral sciences. New York: Academic Press, 1977.

Cohen, J. Eta-squared and partial eta-squared in fixed factor ANOVA designs. Educational and Psychological Measurement, 1973, 33, 107-112.

Conover, W. J. Practical nonparametric statistics. New York: Wiley, 1971.

Coxhead, P. Measuring the relationship between two sets of variables. British Journal of Mathematical and Statistical Psychology, 1974, 27, 205-212.

Cramer, E. M., & Nicewander, W. A. Some symmetric, invariant measures of multivariate association. Psychometrika, 1979, 44, 43-54.

Green, P. E. Mathematical tools for applied multivariate analysis. New York: Academic Press, 1976.

Harris, R. J. A primer of multivariate statistics. New York: Academic Press, 1975.

Harris, R. J. The invalidity of partitioned -U tests in canonical correlation and multivariate analysis of variance. Multivariate Behavioral Research, 1976, 11, 353-365.

Hsu, P. L. On the generalized analysis of variance. Biometrika, 1940, 31, 221-237.

Huberty, C. J. Multivariate indices of strength of association. Multivariate Behavioral Research, 1972, 7, 523-526.

Huberty, C. J., & Mourad, S. A. Estimation in multiple correlation/prediction. Educational and Psychological Measurement, 1980, 40, 101-112.

Keren, G., & Lewis, C. Partial omega squared for ANOVA designs. Educational and Psychological Measurement, 1979, 39, 119-128.

Kerlinger, F. N., & Pedhazur, E. J. Multiple regression in behavioral research. New York: Holt, Rinehart and Winston, 1973.

Knapp, T. R. Canonical correlation analysis: A general parametric significance-testing system. Psychological Bulletin, 1978, 85, 410-416.

Mardia, K. V., Kent, J. T., & Bibby, J. M. Multivariate analysis. New York: Academic Press, 1979.

Maxwell, S. E., Camp, C. J., & Arvey, R. D. Measures of strength of association: A comparative examination. Journal of Applied Psychology, 1981, 66, 525-534.

Mendoza, J. L., Markos, V. H., & Gonter, R. A new perspective on sequential testing procedures in canonical analysis: A Monte Carlo evaluation. Multivariate Behavioral Research, 1978, 13, 371-382.

Morrison, D. F. Multivariate statistical methods. New York: McGraw-Hill, 1976.

Olson, C. L. On choosing a test statistic in multivariate analysis of variance. Psychological Bulletin, 1976, 83, 579-586.

Sachdeva, D. Estimating strength of relationship in multivariate analysis of variance. Educational and Psychological Measurement, 1973, 33, 627-631.

SAS Institute, Inc. SAS 79.5 changes and enhancements. SAS Technical Report P-115. Cary, North Carolina, February 1981.

Serlin, R. C. A multivariate measure of association based on the Pillai-Bartlett criterion. Paper presented at a meeting of the American Educational Research Association, Los Angeles, April 1981.

Shaffer, J. P., & Gillo, M. W. A multivariate extension of the correlation ratio. Educational and Psychological Measurement, 1974, 34, 521-524.

Smith, I. L. The eta coefficient in MANOVA. Multivariate Behavioral Research, 1972, 7, 361-372.

Srikantan, K. S. Canonical association between nominal measurements. Journal of the American Statistical Association, 1970, 65, 284-292.

Stevens, J. P. Global measures of association in multivariate analysis of variance. Multivariate Behavioral Research, 1972, 7, 373-378.

Tatsuoka, M. M. Discriminant analysis. Champaign, Ill.: Institute for Personality & Ability Testing, 1970.

Tatsuoka, M. M. Multivariate analysis. New York: Wiley, 1971.

Tatsuoka, M. M. An examination of the statistical properties of a multivariate measure of strength of relationship (Final Report). Project No. 2-E-020, Grant No. OEG-5-72-0027(509), HEW, December 1973.

Timm, N. H. Multivariate analysis. Monterey, Calif.: Brooks/Cole, 1975.

Wilks, S. S. Certain generalizations in the analysis of variance. Biometrika, 1932, 24, 471-494.

Footnote

[1]Another situation where this caveat applies is in nonorthogonal analysis of variance.